Objectives

Your main task is to cluster the countries and then present your solution and recommendations to the CEO using a PPT. The following approach is suggested :

- Start off with the necessary data inspection and EDA tasks suitable for this dataset - data cleaning, univariate analysis, bivariate analysis etc.
- **Outlier Analysis**: You must perform the Outlier Analysis on the dataset. However, you do have the flexibility of not removing the outliers if it suits the business needs or a lot of countries are getting removed. Hence, all you need to do is find the outliers in the dataset, and then choose whether to keep them or remove them depending on the results you get.
- Try both K-means and Hierarchical clustering(both single and complete linkage) on this dataset to create the clusters. [Note that both the methods may not produce identical results and you might have to choose one of them for the final list of countries.]
- Analyse the clusters and identify the ones which are in dire need of aid. You can analyse the clusters by comparing how these three variables - [gdpp, child_mort and income] vary for each cluster of countries to recognise and differentiate the clusters of developed countries from the clusters of under-developed countries.
- Also, you need to perform visualisations on the clusters that have been formed. You can do this by choosing any two of the three variables mentioned above on the X-Y axes and plotting a scatter plot of all the countries and differentiating the clusters. Make sure you create visualisations for all the three pairs. You can also choose other types of plots like boxplots, etc.
- Both K-means and Hierarchical may give different results because of previous analysis (whether you chose to keep or remove the outliers, how many clusters you chose, etc.) Hence, there might be some subjectivity in the final number of countries that you think should be reported back to the CEO since they depend upon the preceding analysis as well. Here, make sure that you report back at least 5 countries which are in direst need of aid from the analysis work that you perform.

Results Expected

1. A well-commented Jupyter notebook containing the Clustering Models(both K-means and Hierarchical Clustering) and the final list of countries.
2. Present the overall approach of the analysis in a presentation
    a. Mention the problem statement and the analysis approach.
    b. Explain the results of Clustering Model briefly.
    c. Include visualisations and summarise the most important results in the presentation.
    d. Make sure that you mention the final list of countries here (Don't just mention the cluster id or cluster name here. Mention the names of all the countries.)

You need to submit the following two components

- Python notebook: Should include detailed comments and should not contain unnecessary pieces of code
- PPT: Make a PPT to present your analysis to the CEO (and thus you should include both the technical and the business aspects). The PPT should be concise, clear, and to the point. Submit the PPT after converting into the PDF format. The visualisations mentioned above must be present in this file.

Evaluation Rubric

Criteria	Meets expectations	Does not meet expectations
Data understanding and EDA (10%)	All data quality checks are performed, and all data quality issues are addressed in the right way (missing value imputation, removing duplicate data and other kinds of data redundancies, etc.). Explanations for data quality issues are clearly mentioned in comments or in the presentation. New metrics are derived if applicable and are used for analysis and modelling.	All quality checks are not done, data quality issues are not addressed correctly to an appropriate level. Dummy variables are not created properly. New metrics are not derived or are not used for analysis. Univariate and bivariate analysis hasn't been performed The data is not converted to a clean format which is suitable for analysis or is not cleaned using commands in

	Univariate and bivariate analysis has been performed The data is converted to a clean format suitable for analysis in Python.	Python.
Model building and evaluation (45%)	For K-means clustering the logic of choosing K is explained using both the technical aspects as well as the problem-specific aspects. K-means algorithm is utilised properly and at least two iterations are done on different K For hierarchical clustering, the choice of the number of clusters is done logically and	For K-means clustering the logic of choosing K isn't explained coherently. K-means algorithm hasn't been utilised properly For hierarchical clustering, the choice of the number of clusters isn't done logically. The results are not at par with the best possible model on

	the final clusters are shown correctly. The results are on par with the best possible model on the dataset. The model is interpreted and explained correctly. The commented code includes a brief explanation of the important variables and the model in simple terms.	the dataset. The model is not interpreted and explained correctly.
Presentation and Recommendations (10%)	The presentation has a clear structure, is not too long, and explains the most important results concisely in simple language. The recommendations to solve the problems are realistic, actionable and coherent with the analysis.	The presentation lacks structure, is too long or does not put emphasis on the important observations. The language used is complicated for business people to understand. The recommendations to solve the problems are either

	If any assumptions are made, they are stated clearly.	unrealistic, non-actionable or incoherent with the analysis. Contains unnecessary details or lacks the important ones. Assumptions made, if any, are not stated clearly.
Conciseness and readability of the code (5%)	The code is concise and syntactically correct. Wherever appropriate, built-in functions and standard libraries are used instead of writing long code (if-else statements, for loops, etc.). Custom functions are used to perform repetitive tasks. The code is readable with appropriately named variables and detailed	Long and complex code used instead of shorter built-in functions. Custom functions are not used to perform repetitive tasks resulting in the same piece of code being repeated multiple times. Code readability is poor because of vaguely named variables or lack of comments wherever necessary.

	comments are written wherever necessary.	